Contribution ID: **111** Type: **Standard 15 min talk**

# Accelerating Reproducible FPGA Machine Learning Research With a Workflow Management Framework

*Thursday 17 October 2024 13:40 (15 minutes)*

High-Level Synthesis (HLS) techniques, coupled with domain-specific translation tools such as HLS4ML, have made the development of FPGA-based Machine Learning (ML) accelerators more accessible than ever before, allowing scientists to develop and test new models on hardware with unprecedented speed. However, these advantages come with significant costs in terms of implementation complexity. The process of taking code written in a high-level language and translating it to a synthesized hardware IP is a long and fraught one, and configuration and workflow choices made at every step along the way will affect the finished product. Properly documenting each of these subtle choices is difficult, and failing to do so can make it difficult or impossible to reproduce the model. Further complicating matters is the question of optimization - the breadth of possible design choices in modern hardware ML systems is vast, and optimizing these systems by hand is an intractable process. Efficient design space exploration methods are essential in these development flows.

Modern tooling often supports mechanisms to improve the efficiency of design-space explorations. Tensorflow and Pytorch, for example both support data streaming APIs, where training and test datasets can be loaded into memory only as needed. These APIs can have significant performance benefits and enable the exploration of model options that would otherwise have been unfeasible to evaluate due to system resource limitations - our preliminary investigations demonstrated peak memory usage improvements of two orders of magnitude on a very large dataset. However, using these mechanisms requires both awareness that they exist and an investment of development time to make use of them.

As a step toward resolving these issues, we introduce a new open-source framework, the Experimental Setup and Optimization System for HLS4ML (ExSeOS-HLS), which aims to enable optimized and reproducible ML model development flows on hardware systems. By centrally managing all of the steps in the design process, from preprocessing raw input data to extracting result metrics from vendor toolchain reports, ExSeOS-HLS can automatically and reproducibly optimize hyperparameters, HLS settings, and even model architectures for a user-defined target metric or combination thereof. Additionally, it can take advantage of tool-specific development optimizations by default, reducing system resource usage and accelerating the research process with minimal effort on the part of the researcher. Experiment configurations can be exported to a single file, which can be sent to collaborators or published online in order to allow exact reproduction of the research workflow. In introducing this system, our goal is to enable collaborative, reproducible workflows for FPGA-based ML acceleration across many scientific application domains.

## Focus areas

**Primary authors:** SHUPING, Alexis (Northwestern University); Prof. OGRENCI, Seda (Northwestern Univesity)

**Presenter:** SHUPING, Alexis (Northwestern University)

**Session Classification:** Contributed talks